Draft 2

# Applying Weighted Levenshtein Distance in Language Family Homeland Identification

Ziting Shen

April 4, 2018

**Abstract**

Several computational models have been proposed to hypothesize the geographical homeland of a language family in a quantitative manner. Aiming at identifying the homeland accurately for the world's language families with limited data, we propose modifications to one of these models, the ASJP model. Specifically, we apply the weighted Levenshtein distance algorithm to look for the most linguistically diverse geographical region as the language family homeland. The model is tested on the Indo-European family, and the results are compared to major linguistic theories about the Indo-European homeland. The model proposes the Balkans as the homeland, which successfully reflects the high linguistic diversity around the Balkans. Although such hypothesis still diverges from the major linguistic theories, it is believed to be reasonably biased due to the biased dataset and the contentious linguistic assumptions of the model.

Your abstract is great!!!

## 1 Introduction

In the field of historical linguistics, identification of the geographical origin of each language family is always a focus of interest, as it provides insights about evolution of the language family and history of the language speakers. However, it is also one of the most complicated historical linguistic problem. Currently, there are 7099 living languages in the world, most of which are studied by a limited number of linguistic specialists. The 7099 languages are classified into 141 language families, some of which consist of a fair amount of languages and have a wide geographical span (Ethnologue, 2017). For example, the Indo-European family has over 200 languages which are spoken from the Northern Europe to India and from America to Siberia. To identify the homelands for a language family without computational tools, the

linguistic specialists need to manually perform the comparative method, i.e. to compare the lexical items, syntactical structures, and phonological attributes between the languages in the family. Therefore, it is practically hard to manually pinpoint the homeland of some language families due to the high demand of expertise, labor and time, which brings about the necessity to automate homeland identification with computational tools.

This thesis provides a thorough review of past computational linguistic models for language family homeland identification, and proposes a modification to one of them, the Automated Similarity Judgment Program (ASJP) model. The ASJP model proposes that the homeland of a language family can be speculated with respect to the linguistic distance and geographical distance between languages in the family (Wichmann et al., 2010b). The linguistic distance is defined as how different the languages are, while the geographical distance is the physical distance between where the languages are spoken. Specifically, the ASJP model calculate the linguistic distance as the naive Levenshtein distance (introduced in section 3.4) between certain lexical items of languages. This thesis proposes to replace the naive Levenshtein distance calculation with the ALINE algorithm, an algorithm based on weighted Levenshtein distance. By applying weights to the Levenshtein distance algorithm, this thesis aims to incorporate more phonological information into linguistic distance calculation and thus to increase the accuracy of the original ASJP model.

The investigation around the computational models for homeland identification are divided into the following seven sections. Section 2 discusses the motivation of modifying the ASJP model to improve the accuracy of automated language family homeland identification. Section 3 provides a general introduction of the language family homeland identification problem and the linguistic approaches to this problem. Section 4 examines past computational models used for homeland identification, including the ASJP model, the ALINE algorithm, and phylogenetic inference. Section 5 presents a modification of the original ASJP model to incorporate more phonological information during measuring linguistic distance between languages. Section 6 exhibits some sample input and output of the modified model. Section 7 applies the modified model on the Indo-European language family and compares the results with the two major Indo-European hypotheses, the Steppe Theory and the Anatolian Hypotheses. Section 8 discusses work that needs to be done in the future and other possible improvements of the ASJP model.

# 2 Motivation

Previous methods of identifying the homeland of a language family include both computational and non-computational approaches. Without the assistance of computational tools, a homeland is usually identified through the comparative method with the assistance of archaeological and biological evidence. With an increasing interest in the historical expansion and evolution of the language families, computational models become a possible complement to the comparative method for language family phylogeny reconstruction and homeland identification. Computational models are relatively objective during data processing and are able to generate insightful results based on information extraction from raw data. Compared to the comparative method, they require less time, labor and expertise, although they also tend to overlook subtle differences between individual language changes and oversimplify the process of language evolution.

Among the existing models, the ASJP model is prominent for its limited data usage and transparent methodology (Wichmann et al., 2010b). For each language, it only requires the phonetic transcriptions of 40 words and the longitude and latitude of where the language is spoken. Such data is easy to collect and does not need much preprocessing, which is consistent with the goal of saving resources by automating language family homeland identification. Additionally, the methodology of the ASJP model is mainly based on the Levenshtein distance algorithm, which is much simpler than those models based on evolutionary biology and machine learning. Such transparent methodology can be readily understood and further modified by linguistic specialists that do not have much computer science background.

Nevertheless, the ASJP model also shows some major weaknesses that demand recognition and fixation. Firstly, the ASJP model assumes that the language homeland appears at the most linguistically diverse area, which is argued to be untenable for some language families (Mallory and Adams, 2006). Secondly, while the difficulty of collecting data for the ASJP model is greatly reduced by requiring phonetic transcriptions of only 40 words per language, such data might fail to reflect syntactical and morphological features of the language. Thirdly, the ASJP model uses the naive Levenshtein distance to calculate the linguistic distance between language, which omits a lot of phonological information when comparing words.

Among the three major weaknesses mentioned above, the third one can be potentially fixed by applying weights to the Levenshtein distance when comparing words in linguistic distance calculation. In the hope of locating language family homelands

more accurately, we combine the ALINE algorithm, a weighted Levenshtein distance algorithm, with the ASJP model to take more phonological nuances into consideration (Kondrak and Hirst, 2002). In ALINE algorithm, each word is represented as a sequence of phonemes. When comparing two words, different weights are assigned to different phoneme operations. The operations include insertion of a phoneme, deletion of a phoneme, substitution of a phoneme for another one or two phonemes, and substitution of two phonemes for another one. The weighted Levenshtein distance will be able to reveal the difference in commonness of different phoneme operations in natural languages.

# 3 Background

## 3.1 Language Family Structure

When historical linguistics examines the changes of languages over time, it groups languages into language families based on their similarity (Millar and Trask, 2015, p. 167). The linguistic similarity is assumed to signify the genetic relationships between the languages, i.e. languages within a language family are assumed to be descendants of the same ancient language, the proto-language of the family. The homeland of the family, the region where the proto-language was spoken, can be identified by analyzing the evolution of the family and the ethnographic information carried in the languages. Therefore, consistently increasing interest in the topic of homeland identification is shown not only in the field of linguistics but also in the fields of archaeology, anthropology, and evolutionary biology.

The homeland location can be hypothesized through examination of the structure of a language family, especially the major divergences of the family. A tree structure is usually used to model the structure of the family, since it is assumed that new languages are constantly created by the divergence of extant languages, and that they will never merge back (Millar and Trask, 2015, p. 169). The resulting tree structure is called a language family tree, which is modeled on the phylogenetic tree between species in evolutionary biology. Because of the parallels, it is called phylogenetic reconstruction or phylogenetic inference. Reconstruction of the phylogeny is valued highly when identifying homelands, because the phylogeny reflects the linguistic closeness between languages and implies the temporal and spatial relationships between them as well. Some computational models take the phylogenetic inference approach and apply computational biological models to reconstruct the language fam-

ily phylogeny and identify the homeland, which will be introduced in more detail in Section 4.1.

## 3.2 Linguistic Similarity

Linguistically, the homeland is hypothesized through examining the tree structure of the family. Some languages in the family resemble more than the others and exclusively share some linguistic features. Fore such shared features, there are several possible explanations (Millar and Trask, 2015, p. 17, p. 169). The shared features might be shared innovations, i.e. these languages originate from a common ancestor that develops these linguistic features and diverges from the rest of the family. Alternatively, the shared features could be shared archaisms, i.e. these features are inherited from the oldest ancestor of the family. While some languages preserve these features, others just drop them. Generally, the shared innovations can suggest genetic closeness between certain languages while the shared archaisms cannot, because every language in the family have the chance to inherit the shared archaisms from the oldest ancestor, but not every language is able to inherit the shared innovations from a particular ancient language.

Another possibility is that the shared features could be borrowings from neighboring language families. Specifically, the features could be borrowed by each language individually or by their common ancestors. They could be borrowed directly from the donor language, or successively through an intermediate language. That being said, the origin and development of each feature should be scrutinized to deduce the relative geographical distribution of the languages which share the common features.

In order to distinguish between shared innovations, shared archaisms, and borrowings, linguists employ the comparative method to perform a feature-by-feature comparison and reconstruct the ancestor of genetically related languages (Millar and Trask, 2015, p. 191-192). Performing the comparative method requires a lot of linguistic expertise, labor, and time, which has been recognized by some linguists and leads into a number of attempts to simplify the identification methodology. Most of these simplification attempt to produce quantitative judgments on the structure and evolution of a language family. Nevertheless, the simplified approaches are widely criticized for overgeneralization and oversimplification of language family evolution. Since the ASJP model also relies on some of the controversial linguistic methods, the following paragraphs will discuss them in detail.

On the basis of the comparative method, one simplification – called lexicostatis-

tics – is to limit the range of items being compared between languages by only comparing a small set of "basic vocabulary". The "basic vocabulary" are words that correspond to a list of basic concepts in daily life, such as body parts, numerals, elements of nature, etc. This list of vocabulary was originally proposed by Swadesh (1950) and later called the Swadesh list. According to Swadesh, the vocabulary in such a list should be resistant to borrowing because of the difficulty of altering the "basic vocabulary" in a language. Therefore, through identifying the shared innovations and comparing their percentage in the Swadesh list, lexicostatistics is able to infer closeness between languages within a language family.

Lexicostatistics is severely criticized for their oversimplification of the comparative method. For the Swadesh list used in lexicostatistics, usually one word is arbitrarily chosen for one meaning (Millar and Trask, 2015, p. 352). However, in reality, the same meaning might be expressed by multiple words whose meanings have nuanced differences. Among these synonymous words, some might be decedents of the same word in the ancient language, while the others have unrelated origins. Without detailed examination, it is possible for a Swadesh list to omit some critical properties of the lexicons and morphology of a language.

## 3.3 Indo-European Homeland

The model proposed by this thesis will be tested on the Indo-European family, a well-studied language family that contains more than 200 languages, most of which are spoken in Europe, America, and India. In that case, this section will briefly introduce the two linguistic theories about the Indo-European homeland which are supported by the majority of linguists.

The two major theories about the Indo-European homeland are respectively called the Steppe Theory and the Anatolian Hypotheses. Shown in Figure 3.1, the Steppe Theory proposes that the Indo-European family emerged out of local communities in the Pontic-Caspian Steppe around 4000 B.C. (Mallory and Adams, 2006, p. 461-462). The family then expanded westwards and eastwards by mobile communities through horses and wheeled vehicles.

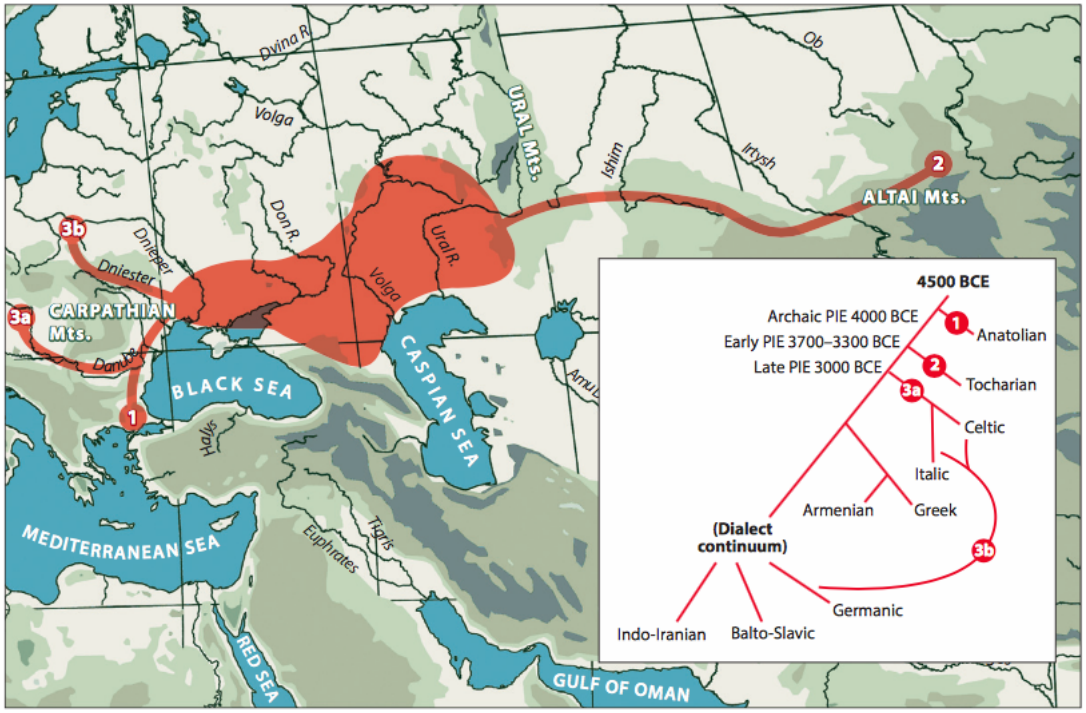



Figure 3.1: (Anthony and Ringe, 2015, p. 209) Map of the Steppe Theory. The red area represents the Pontic-Caspian Steppe, the Indo-European homeland hypothesized by the Steppe Theory. The curves connected to the red areas shows the direction of migration at about 4200 BC (1), 3300 BC (2), and 3000 BC (3a and 3b). The tree shows the major divergences of the Indo-European family and their times hypothesized by the Steppe Theory.

Shown in Figure 3.2, the Anatolian Hypothesis argues that the Indo-European family originated in Anatolia around 7000 B.C. and gradually spread to peripheral areas by the earliest farming communities (Mallory and Adams, 2006, p. 460-461). The biggest dispute about this hypothesis is that some technological items attested across Indo-European branches, e.g.wheeled vehicles and the plough, were invented later than 4500 B.C. in the Late Neolithic or Early Bronze Age.

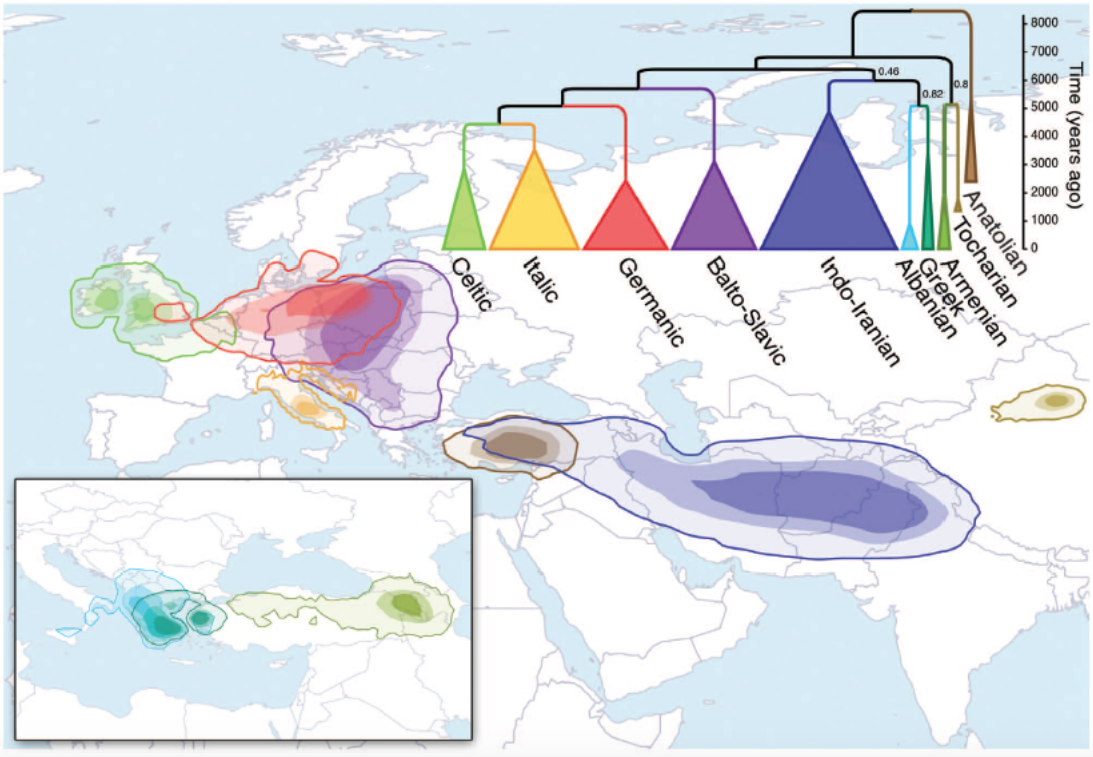



Figure 3.2: (Bouckaert et al., 2012, p. 959) Map of the Anatolia Hypothesis. The tree shows the major divergences of the Indo-European family and their times hypothesized by the Anatolian Hypothesis. The colored areas are the inferred locations of each divergence. The brown area is the Anatolia, the hypothesized Indo-European homeland according to the Anatolian Hypothesis.

## 3.4 Levenshtein Distance

As the model proposed by this thesis will replace naive Levenshtein distance calculation with the ALINE algorithm to compute weighted Levenshtein distance between words, this section will introduce both the naive(non-weighted) and weighted Levenshtein distance algorithm.

The Levenshtein distance algorithm was initially proposed by Levenshtein (1966) as the minimal number of string operations to transform one string to another. Initially, the string operations only include insertions and deletions of characters, thus substitution of characters would be counted count as two operations. Later, substitution is added as another string operation, which means it will only be counted as one operation.

To calculate the Levenshtein distance, Wagner and Fischer (1974) propose a dynamic programming algorithm to construct a matrix of the Levenshtein distance of

substrings of the two strings compared and obtain the Levenshtein distance between the two complete strings after filling the matrix. Table 3.1 shows a Levenshtein distance matrix between "sneak" and "snake". After filling the matrix, the value in the right bottom cell will be taken as the Levenshtein distance between the two words.

| | | s | n | e | a | k |
|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 |
| s | 1 | 0 | 1 | 2 | 3 | 4 |
| n | 2 | 1 | 0 | 1 | 2 | 3 |
| a | 3 | 2 | 1 | 1 | 1 | 2 |
| k | 4 | 3 | 2 | 2 | 2 | 1 |
| e | 5 | 4 | 3 | 2 | 3 | 2 |

Table 3.1: An example of Levenshtein distance matrix between "sneak" and "snake." The Levenshtein distance between them is 2, as shown in the right bottom cell.

For the cell $L(i, j)$ at $i$th row and $j$th column in a Levenshtein distance matrix, Wagner and Fischer (1974) state that the cell should be filled using the following recurrence relation, where $\text{word1}_i$ is the $i$th character in word1 and $\text{word2}_j$ is the $j$th character in word2. The value in cell $L(i, j)$ signifies the least number of operations to transform the prefix of word1 of length $i$ to the prefix of word2 of length $j$. For example, cell $L(4, 3)$ has value 2, which means that it takes at least 2 operations, i.e. a substitution and an insertion, to transform "snak" to "sne."

$$L(i,j) = \begin{cases} \infty & \text{if } i < 0 \text{ or } j < 0 \\ 0 & \text{if } i = 0 \text{ and } j = 0 \\ L(i-1, j-1) & \text{if } \text{word1}_i = \text{word2}_j \\ \min(L(i-1,j)+1, & \\ \quad L(i,j-1)+1, & \text{otherwise} \\ \quad L(i-1,j-1)+1) & \end{cases}$$

In the recurrence relation above, length $i$ and $j$ cannot be negative, so $L(i, j)$ will be $\infty$ when one of them is negative. When both $i$ and $j$ are 0, $L(i, j)$ is 0, because it takes no operation to transform an empty string to another empty string. When $\text{word1}_i$ is the same as $\text{word2}_j$, no operation is needed to transform $\text{word1}_i$ to $\text{word2}_j$, therefore $L(i, j)$ is equivalent to $L(i-1, j-1)$. When $\text{word1}_i$ and $\text{word2}_j$ are different, it must take one operation to substitute between the two characters, or add one of them to the shorter prefixes.

When the Levenshtein distance algorithm is initially proposed, all the string operations are equally weighted, i.e. have weight 1. In real-world applications, they can also be unequally weighted based on the category of string operations or characters involved in the operations.

In the previous example of "sneak" and "snake," the Levenshtein distance are calculated using the orthographic forms, i.e. spellings, of the two words. In the field of linguistics, it is usually desirable to compare the pronunciation of words rather than the spelling, thus the Levenshtein distance can also be calculated using their phonological forms. Table 3.2 shows a Levenshtein distance matrix between the phonological forms of "sneak" and "snake" in the International Phonetic Alphabet (IPA), /sniːk/ and /sneɪk/.

| | | s | n | iː | k |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 |
| s | 1 | 0 | 1 | 2 | 3 |
| n | 2 | 1 | 0 | 1 | 2 |
| e | 3 | 2 | 1 | 1 | 2 |
| ɪ | 4 | 3 | 2 | 2 | 2 |
| k | 5 | 4 | 3 | 3 | 3 |

Table 3.2: An example of the phonological Levenshtein distance matrix between "sneak" /sniːk/ and "snake" /sneɪk/. The phonological Levenshtein distance between the two words is 3, as shown in the right bottom cell.

# 4 Past Computational Models

On the subject of homeland identification, computational historical linguistics provides models that can be used as complements to the comparative method. Instead of performing a qualitative comparison between the languages manually, the computational methods incorporate quantitative measurements into the traditional comparative method. They can hypothesize the homeland of a language family by quantifying differences between languages, conjecturing their subgrouping, identifying their geographical distribution, and inferring the historical expansion of the family.

In past computational models, quantification of language differences can be done with or without identifying vocabulary originated from the same ancient language. Such vocabulary are called "cognates" in the field of linguistics. Up to now, accurate identification of cognates still requires human annotation, which brings higher accuracy for the computational models yet also increases the difficulty of collecting data.

For those language families that are not well studied, such human-annotated data can be fairly expensive to generate. As one of the desirable feature of computational models are their economization of resources, models that require finely processed data will defeat their own purpose.

If cognates are not identified, then the language differences will be computed as the phonological distance between languages. Models following this approach do not limit the objects of phonological comparison to cognates only, but directly calculate the phonological distances between semantically equivalent items and hypothesizes the subgrouping of languages based on their phonetic similarity.

As argued by Downey et al. (2008) and Heggarty (2000), skirting the subject of cognate identification can be both the strength and weakness of some approaches. Traditionally, language family trees are constructed through examining the cognates only, thus presumably, not distinguishing non-cognates from cognates will lead to inaccurate homeland identification results. However, not eliminating the non-cognates from the data preserves more data that contains more phonological nuances, which might be informative in measuring the linguistic distance between languages. Additionally, models of this approach are more scalable in terms of data availability and computational efficiency. For language families that are not well-studied, these models can usually be readily applied without much modification and training.

## 4.1 Phylogenetic Inference

Phylogenetic inference is a computational homeland identification model that requires cognate identification. Compared to other computational models, phylogenetic inference involves a relatively complex data preprocessing process, yet it yields the most plausible results. It originates from an evolutionary biological model for reconstruction of the evolutionary tree of species. Both species and languages inherit features from ancestors and generate mutations that differentiate them from the previous generation. For species, the mutations that promote exploitation and utilization of resources are more likely to be passed to subsequent generations. Similarly, for languages, the mutations that carry social capital are more likely to be passed on. Since it is feasible to model the evolutionary process computationally, phylogenetic inference is adopted as one of the computational linguistic models for reconstructing the history of a language family (Gray and Atkinson, 2003).

Phylogenetic inference algorithm models the evolution of a language family by examining the traits of current and extinct languages. Each language is represented

as a character list: a list of the language's lexical, syntactical, morphological and phonological changes in the history. For the lexical changes, the character list distinguishes between cognates and borrowings, i.e. between vocabulary inherited from a common ancestor or borrowed from another language. For the phonological changes, they can be compiled manually or extracted automatically from a list of cognates. Manually compiled cognate lists are usually used, as they are more accurate than the automatically generated ones. The character lists are then used to estimate the language family tree and major divergence time under models of evolution. During estimation, probabilistic models can be incorporated to mimic the uncertainty in the evolutionary process, and rate-smoothing algorithms can be added to model the variation of rate of change for each language. Specifically, rate variation is allowed between different branches of a language family, yet the change of rate between branches will be penalized using the penalized-likelihood model. By doing so, the phylogenetic inference algorithms avoid some of the oversimplification and problematic assumptions of lexicostatistics and glottochronology.

Following the approach of phylogenetic inference, Gray and Atkinson (2003) construct the family tree and divergence times of the Indo-European languages. In their results, the family tree is consistent with the traditional Indo-European groupings. The divergence times match those proposed by the Anatolian Hypothesis, one of the major theories about the Proto-Indo-European homeland. Based on the results of Gray and Atkinson (2003), Bouckaert et al. (2012) estimate that the homeland of the Indo-European family is highly likely to be within the Anatolian peninsula, which also strongly supports the Anatolian Hypothesis.

Undeniably, the phylogenetic inference approach is powerful and offers reasonable results. However, its demand for finely processed data – such as manually created character lists or manually annotated cognate lists – to some extent defeats its own purpose as a quantitative computational model. Generally, computational models are valued not only for their objectivity and perspicacity during information extraction from raw data, but also for their economization of resources used in data processing. In that sense, the phylogenetic inference approach is not as beneficial as other approaches that require less preprocessing of data.

## 4.2 The ASJP Model

The ASJP (Automated Similarity Judgement Program) model is one of the models that do not require cognate identification. As the only model that has been applied

to more than half of the world's language families, it pinpoints the homelands of 256 Ethnologue language families comprised of 4664 total languages (Wichmann et al., 2010b). Its results are fairly contentious, mainly due to its two main assumptions. The first assumption it makes is the center of gravity principle, which specifies that the homeland will appear at the most linguistically diverse region within the language family. The underlying logic of the principle is that as an ancestor language evolves, it gradually develops to multiple languages that have different features. The multiple languages will become more and more linguistically distant from each other and finally evolve into their own descendant languages, leading to the divergence of subgroups within the family. Provided the rate of change is constant enough in a language family, big differences between language subgroups will necessarily indicate an early divergence. Therefore, the maximal diversity within a language family should appear near the earliest divergence, which happens at the homeland of the language family.

As discussed in section 3, the center of gravity principle is problematic, because the linguistic diversity does not only come from divergences within a language family, but can also be brought by contact with neighboring languages (Mallory and Adams, 2006, p. 445-446). The rate of change within a language family is also not constant so that larger diversity does not always indicate earlier divergences (Bergsland and Vogt, 1962). Moreover, after migration for thousands of years, several branches of the same language family might coincidentally end up in the same geographical regions and increase the linguistic diversity of the region, which is exemplified by the Indo-European subfamilies in the Balkans (Mallory and Adams, 2006, p. 451-452).

The second contentious assumption of the ASJP model is that the phonetic distance between languages can be measured through a small set of "basic vocabulary", such as the Swadesh list. Among the standard 100-item Swadesh list, ASJP uses a subset of 40 items to measure the phonetic distance between languages (Holman et al., 2008). The semantic meanings of the 40 items are listed as below (Bakker et al., 2009):

> I, You, We, One, Two, Person, Fish, Dog, Louse, Tree, Leaf, Skin, Blood, Bone, Horn, Ear, Eye, Nose, Tooth, Tongue, Knee, Hand, Breast, Liver, Drink, See, Hear, Die, Come, Sun, Star, Water, Stone, Fire, Path, Mountain, Night, Full, New, Name

As discussed in the previous section, the vocabulary selected for each meaning in the Swadesh list might significantly change the results of the model, because some meaning can correspond to multiple words whose meanings have nuanced differences.

Among the multiple words, some of them might resemble words with the same meaning in another language, while the others might not. Moreover, lexical item borrowing is expected not to appear in the "basic vocabulary," so that the linguistic diversity calculated from them should all result from divergences within the language family, which can best demonstrate the closeness between languages. Nevertheless, in practice lexical item borrowing still exists in the "basic vocabulary." For example, for the 40 meanings shown above, the English words "person" and "mountain" are borrowed from Old French (Onions, 1996).

Despite these problematic assumptions which oversimplify the process of language evolution, the ASJP model is indeed simple and powerful. Based on phonetic transcriptions of a moderately limited number of words, it identifies the homeland for a language family in the following manner.

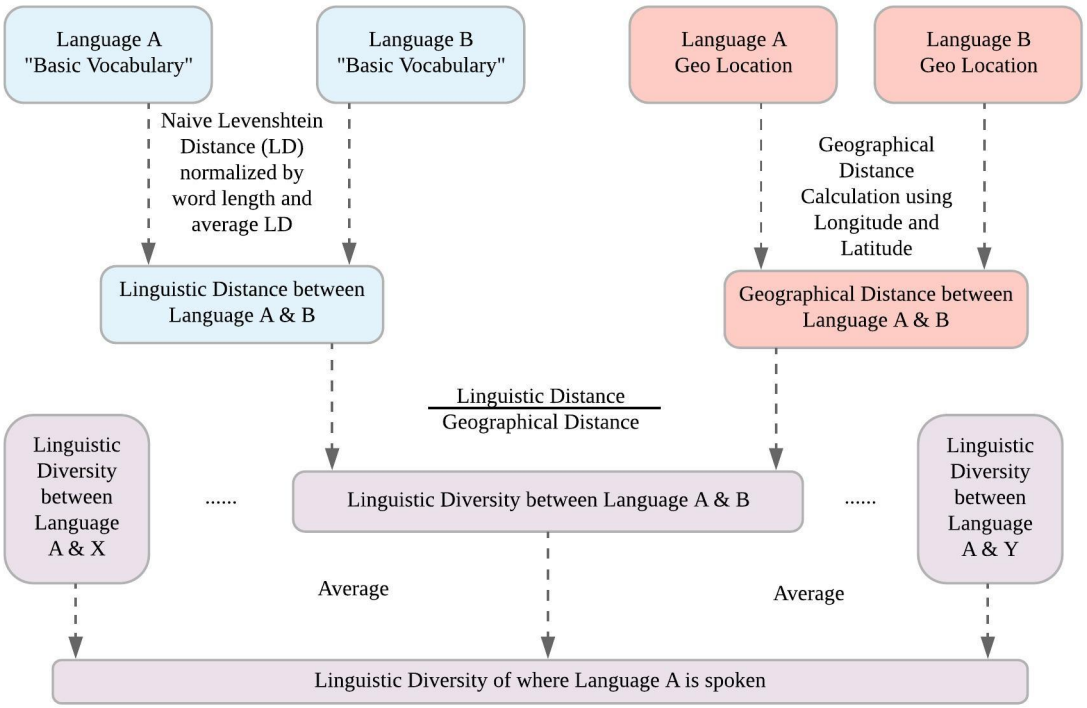



Figure 4.1: Block diagram of the ASJP model.

1. **Calculate the LD (Levenshtein distance).** For any two languages in the family, the ASJP model calculates the LD for each pair of semantically equivalent words in their 40-item Swadesh lists (or equivalent) by counting the least number of operations to transform one word to another through inserting, delet-

ing, or substituting phonemes (Levenshtein, 1966).

2. **Calculate the LDN (Levenshtein Distance Normalized).** The LDN is calculated through normalizing the LD by the length of the longer word between the two semantically equivalent words.

3. **Calculate the LDND (Levenshtein Distance Normalized Divided).** The LDND is computed by normalizing the LDN by the average LDN of all pairs of semantically equivalent words between the two languages.

4. **Obtain the linguistic distance between two languages.** The linguistic distance between two languages is the average LDND of all pairs of semantically equivalent words between them.

5. **Obtain the geographical distance between two languages.** The geographical distance between two languages is the distance between two points that represent where each language is spoken. The points are expressed in longitude and latitude.

6. **Measure the linguistic diversity brought by two languages.** The linguistic diversity brought by two languages to the region between where they are spoken is the ratio between their linguistic and geographical distances.

7. **Measure the linguistic diversity of homeland candidates.** The ASJP model assumes the homeland is located where a current language is spoken. For each current language, the model averages the diversity between it and every other language in the family and considers the averaged diversity as the diversity of where the language is spoken. Notably, the model does not pinpoint the homeland at the center of gravity of the linguistic diversity, because the center of gravity might be in water or depopulated zones.

8. **Pinpoint the homeland.** The diversity of all homeland candidates are ranked from the highest to the lowest. The more linguistically diverse a geographical location is, the more likely it is to be the homeland.

Figure 4.2 shows the homelands of Eurasian language families hypothesized by the ASJP model.

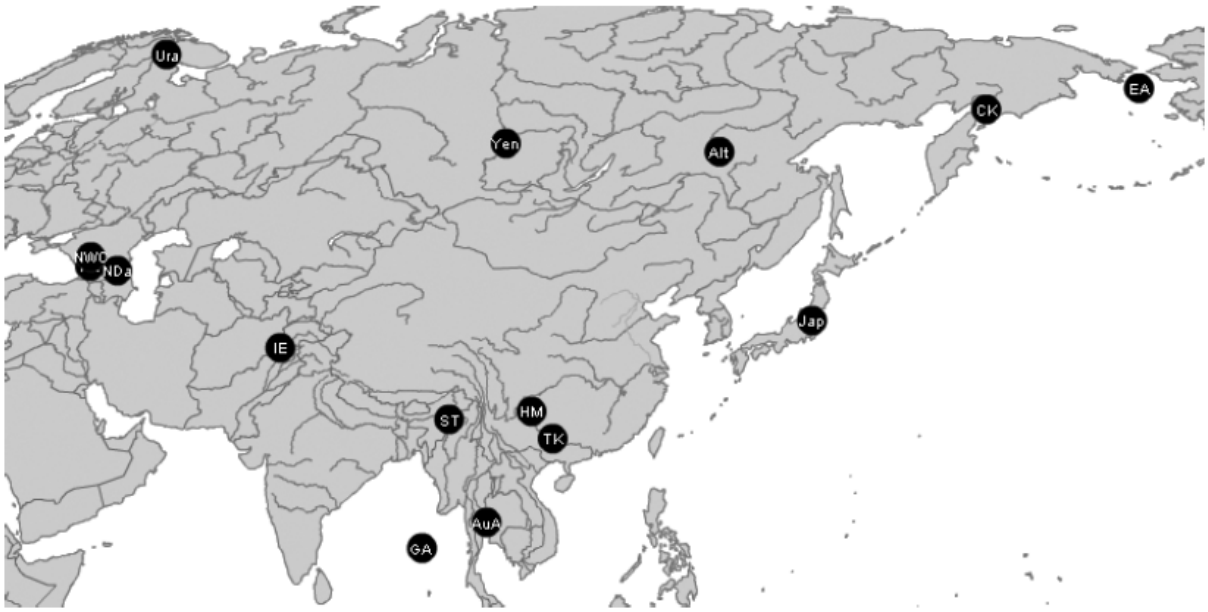



Figure 4.2: (Wichmann et al., 2010b, p. 259) Homelands of Eurasian language families hypothesized by the ASJP model.

Legend: *Alt*: Altaic; *AuA*: Austro-Asiatic; *CK*: Chukotko-Kamchatkan; *EA*: Eskimo-Aleut; *GA*: Great Andamanese; *HM*: Hmong-Mien; *IE*: Indo-European; *Jap*: Japanese; *Krt*: Kartvelian; *NDa*: Nakh-Daghestanian; *NWC*: Northwest Caucasian; *ST*: Sino-Tibetan; *TK*: Tai-Kadai; *Ura*: Uralic; *Yen*: Yeniseian. Note: *NWC* is superimposed on *Krt*.

Among the results of the ASJP model, the hypothesized Indo-European homeland is highly controversial and concerns a lot of Indo-European specialists. The ASJP model examines 40-item Swadesh lists of 220 Indo-European languages that cover all currently existing major branches of the Indo-European family. In other words, all data for languages that have been extinct for longer than three centuries is eliminated, and some major branches, like the Tocharian and Anatolian subfamilies, are not represented in the data. Through comparing the linguistic diversity on such data, the ASJP model ranks the Indo-European homeland candidate locations in the order of Northern Pakistan, the Balkans, and Eastern Anatolia. Such conjectures diverge greatly from the two major theories about the Indo-European homeland: the Steppe Theory and the Anatolian Hypothesis (Mallory and Adams, 2006, p. 460-463). Figure 4.3 compares the Indo-European homeland hypothesized by the ASJP model with the Steppe Theory and the Anatolian Hypothesis.

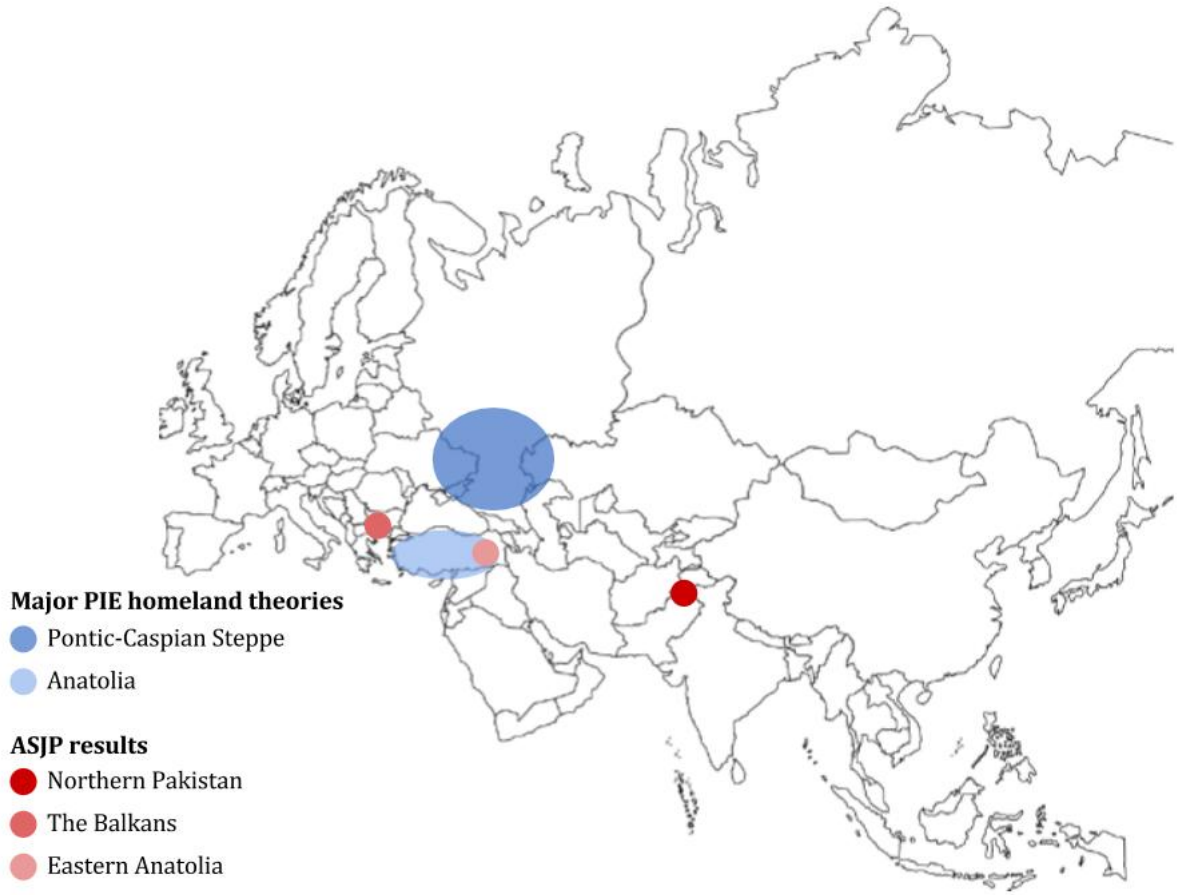



Figure 4.3: The hypothesized Indo-European homeland of ASJP results and major PIE (Proto-Indo-European) homeland theories.

To some extent, the divergence of results of the ASJP model from the conclusion of expets is due to its application of the center of diversity principle. As foreseen by Mallory and Adams (2006), the principle will bias the hypothesized Proto-Indo-European homeland to the linguistically diverse Balkans. Moreover, the ASJP only considers current languages or those that have gone extinct within the last three centuries. Its omission of some important extinct branches, like the Anatolian and Tocharian subfamilies, and insufficiency of data for some current languages might also skew the results to some extent.

Another potential cause of the divergence might be from the encodings of the data. The ASJP model represents the phonetic transcription of the 40-item Swadesh list by ASJPcode shown below.

Consonants

| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plosive | p b | | t d | | | | T | k g | q G | | 7 |
| Nasal | m | | 4 | n | | | 5 | N | | | |
| Trill | | | r | | | | | | r | | |
| Fricative | p b | f v | 8 | s z | S Z | | | x | X | | h |
| Approximant | | | r | | | | y | Bilabial-velar: w | | | |
| Lateral approximant | | | l | | | | | | | | |
| Affricate | Alveolar: c Palato-alveolar: C | | | | | | | | | | |
| Clicks | ! | | | | | | | | | | |

Vowels

| | Front | Central | Back |
|---|---|---|---|
| Close | i | 3 | u |
| Close-mid | e | 3 | o |
| Open-mid | E | 3 | o |
| Open | E | a | o |

Other Symbols

| | | |
|---|---|---|
| * | Nasalized vowel | a* |
| ∼ | Two juxtaposed consonants | tw∼ |
| $ | Three juxtaposed consonants | ndy$ |
| ” | Glottalized consonant | k” |

Table 4.1: ASJPcode, the encoding system of the ASJP model, is consisted of which consists of 41 symbols representing 7 vowels and 34 consonants (Brown et al., 2008).

| Meaning | English | Standard German |
|---|---|---|
| ‘I’ | Ei | ix |
| ‘you’ | yu | du |
| ‘we’ | wi | vir |
| ‘one’ | w3n | ains |
| ‘person’ | pers3n | mEnS |

Table 4.2: Examples of ASJP data.

As shown in Table 4.1, some symbols in ASJPcode cover a broad range of sounds. The lengths, accents or tones of vowels are not represented, either. Additionally, certain complex syllable nuclei are reduced to simple syllable nuclei: CVhC, CV7C, CVxC, CvXC, and CVyC are all reduced to CVC (where C is a consonant and V is a vowel). As a results, the ASJPcode oversimplifies phonetic transcriptions of a lot of words, which results in the failure of representing certain phonetic differences in Levenshtein distance calculation.

In conclusion, most shortcomings of the ASJP model come from its attempts to gain computational simplicity by employing disputable assumptions. However, as the only model which has been applied on most of the world’s language families, it has exhibited its unique advantage in economicality and scalability. The transparency

and simplicity embedded in its methodology also leave considerable space for further modifications and experiments.

### 4.2.1 ALINE Algorithm

One of the major algorithm used by the ASJP model to compute the linguistic distance between languages is the naive Levenshtein distance algorithm, which is introduced in section 3.4. As the word transformation operations are simply categorized into insertion, deletion, and substitution of phonemes, it suffers from the inability to distinguish the relative distance between phonemes. For instance, the cost of substituting /p/ for /b/ and substituting /n/ for /b/ are the same in naive Levenshtein distance, which fails to reflect that the voiced bilabial stop /b/ is more distinct from the alveolar nasal /n/ than the voiceless bilabial stop /p/.

Compared to the naive Levenshtein distance algorithm, the ALINE algorithm incorporates more phonological knowledge into phonetic similarity measurements by performing comparisons on the features rather than whole phonemes (Kondrak, 2000; Kondrak and Hirst, 2002; Huff and Lonsdale, 2011). Specifically, it preserves the relative distance between phonemes through detailed examination of word transformation operations. In the ALINE algorithm, each word is represented as a sequence of phonemes, while each phoneme corresponds to a set of multivalued features. The ALINE distance between two semantically equivalent words is measured in the following manner.

Firstly, the ALINE algorithm measures the ALINE similarity between the two words. Essentially, the ALINE similarity indicates how similar the two words are. It is measured through examining all possible combinations of operations on phonemes to convert one word to another. Each operation can be insertion, deletion, or substitution of a phoneme, breaking a phoneme into two, or coalescence of two phonemes into one. Usually two phonemes involved in a operation have different feature values for some features. The similarity involved in a certain phoneme operation is calculated as following: the product of the feature value difference and the feature salience is summed up for all features and then subtracted from the cost of the operation. The ALINE similarity is the maximized aggregate of the similarity involved in all phoneme operations. For two words, a high ALINE similarity score between them signifies they are similar, while a low score signifies they are different.

$$\text{ALINE similarity} = max(\sum_{\text{operation}} (\text{operation cost} - \sum_{\text{feature}} (|\text{word1 feature value} - \text{word2 feature value}| * \text{feature salience})))$$

The feature values, feature saliences and operation costs are shown in Table 4.3.

| Feature Name | Value Name | Value |
|---|---|---|
| Place | Bilabial | 100 |
| | Labiodental | 95 |
| | Dental | 90 |
| | Alveolar | 85 |
| | Retroflex | 80 |
| | Palato-alveolar | 75 |
| | Palatal | 70 |
| | Velar | 60 |
| | Uvular | 50 |
| | Pharyngeal | 30 |
| | Glottal | 10 |
| Manner | Stop | 100 |
| | Affricate | 90 |
| | Fricative | 80 |
| | Approximant | 60 |
| | Trill | 50 |
| | Vowel | 40 |
| | High-vowel | 40 |
| | Mid-vowel | 20 |
| | Low-vowel | 0 |
| High | High | 100 |
| | Mid | 50 |
| | Low | 0 |
| Back | Front | 100 |
| | Central | 50 |
| | Back | 0 |

| Feature Name | Salience |
|---|---|
| Syllabic * | 5 |
| Place | 40 |
| Voice * | 10 |
| Nasal * | 10 |
| Lateral * | 10 |
| Aspirated * | 5 |
| High | 5 |
| Back | 5 |
| Manner | 50 |
| Retroflex * | 10 |
| Long * | 1 |
| Round * | 5 |

| Operation | Cost |
|---|---|
| Skip | -1000 |
| Substitution | 3500 |
| Expansion | 4500 |
| Vowel[a] | 1000 |

[a]If one of the phonemes involved in a substitution/expansion operation is a vowel, then the original operation cost needs to decrease by 1000 (the vowel operation cost). If both phonemes are vowels, the original operation cost needs to decrease by 2000.

Table 4.3: (Huff and Lonsdale, 2011) On the left are multivalued features and their values (Kondrak and Hirst, 2002). On the top right are the feature saliences, including basic features and addable features marked by asterisk. On the bottom right are the costs for phoneme operations. The feature values, feature saliences, and operation costs are arbitrarily chosen based on *a priori* linguistic knowledge.

Table 4.4 shows an example of computing the ALINE similarity between the words *cat* and *car*. The vowel in *cat* is marked as front in the ALINE encoding (Huff and Lonsdale, 2011). The value in each cell is the similarity between the truncated phoneme sequences of the two words. For instance, 4750 is the similarity between 'ca' and 'caF'.

| | | c | a | r |
|---|---|---|---|---|
| | 0 | 0 | 0 | 0 |
| c | 0 | 3500 | 2500 | 1500 |
| aF | 0 | 2500 | 4750 | 3750 |
| t | 0 | 3000 | 3750 | 4050 |

Table 4.4: (Huff and Lonsdale, 2011) A sample ALINE distance matrix for computing the difference between *cat* and *car*. The largest value in the matrix is the overall similarity between th two words.

After the ALINE similarity between a pair of words is calculated, it will be normalized with respect to the self-similarity of the two words and converted to the distance between them using the following equation (Downey et al., 2008). The self-similarity of a word is the similarity calculated between the word and itself. For words of different lengths that contain different phonemes, their self-similarity would be different.

$$d_{\text{ALINE}} = 1 - \frac{2 * \text{similarity(word1, word2)}}{\text{similarity(word1, word1)} + \text{similarity(word2, word2)}}$$

Finally, the ALINE distance between two languages is obtained through averaging the ALINE distance for all the pairs of semantically equivalent words between the two languages.

Notably, the ALINE algorithm does not directly calculate the distance, because the same phonological difference should be weighted less in longer, more complicated phonetic sequences. For instance, /bat/ and /bot/ are more similar and less distant than /a/ and /o/.

Compared to the naive Levenshtein distance, the ALINE algorithm provides an effective option to trade some computational efficiency for a more linguistically informed model. While still using raw data, it successfully extracts more phonetic characteristics of a language which can be used in evolution reconstruction and homeland identification of language families.

One of the applications of the ALINE algorithm in historical linguistics is to construct language phylogenies with the assistance of clustering methods. For example, Downey et al. (2008) compared the language trees constructed for 18 languages on Sumba, east Indonesia using the traditional comparative method and the ALINE algorithm with the clustering method UPGMA (Unweighted Pair Group Method with Arithmetic mean). 87.9% nodes and edges in the ALINE tree match with the ones in the tree constructed by the comparative method. Downey et al. summarized that the ALINE algorithm performed well with distance-based tree-building algorithms

like UPGMA, yet they also recognized that the ALINE algorithm was unable to distinguish if some lexical/phonological change is caused by language contact or genetic relationships.

## 4.3 Pros and Cons of the Past Models

Table 4.5 and 4.6 overview the models discussed to this point and summarize their pros and cons.

| Model | Pros and Cons |
|---|---|
| Naive Levenshtein Distance | Pros:<br>1. It has transparent methodology.<br>2. It can be run in very limited time, thus it is able to process more data.<br>Cons:<br>1. It categorizes word transformation operations as phoneme insertion, deletion, and substitution, which omits the relative distance between phonemes. |
| ALINE algorithm | Pros:<br>1. It examines detailed phonological information when measuring the linguistic similarity.<br>2. It is more informative than naive Levenshtein distance.<br>Cons:<br>1. It takes longer to run the algorithm. |

Table 4.5: Pros and cons of the models of linguistic distance measurement.

| Approach | Pros and Cons |
|---|---|
| Phylogenetic inference | Pros:<br>1. It has highly plausible results.<br>2. It considers lexical, syntactical, morphological, and phonological changes.<br>Cons:<br>1. It requires complex manual preprocessing of data.<br>2. It is hard to apply to not well-studied language families. |
| ASJP model | Pros:<br>1. It requires simple data preprocessing.<br>2. It has transparent, modifiable methodology.<br>3. It is equally applicable for different language families.<br>Cons:<br>1. It uses the dubious assumption that the homeland is located in the most linguistically diverse area.<br>2. It assumes languages change at a "constant enough" rate, which might not be the case in practice.<br>3. It assumes semantically equivalent words in the Swadesh list are cognates, although they might actually have unrelated origins. |
| Modified ASJP model proposed in this thesis | Pros:<br>1. It requires simple yet linguistically informative raw data.<br>2. It has transparent methodology.<br>3. It is universally applicable to the world's language families.<br>4. It measures the linguistic distance between languages more accurately and thus performs better than the original ASJP model.<br>Cons:<br>The same as ASJP model. |

Table 4.6: Pros and cons of the approaches to identifying language family homeland.

# 5 Modification of the ASJP Model

This thesis presents a modification of the ASJP model to incorporate more phonological information when calculating the linguistic distance between languages. In

the original ASJP model, the linguistic distance is measured as the naive Levenshtein distance between semantically equivalent words in coarsely-transcribed phonological forms. It categorizes word transformation operations into insertion, deletion, and substitution of phonemes and ignores that some phonemes are more phonologically similar than others. Aiming at representing the variation of relative distances between phonemes, this thesis proposes to quantitatively measure the phonological distances through the modified ALINE algorithm in linguistic distance measurement component of the ASJP model. Compared to the phylogenetic inference approach, the ASJP model and ALINE algorithm have transparent methodology and require simple preprocessing of raw data. Built based on the ASJP model and ALINE algorithm, the modified model is also readily comprehensible, universally applicable, and highly adaptable for further modifications. Figure 5.1 is a block diagram of the modified ASJP model, where the modifications are marked as red.

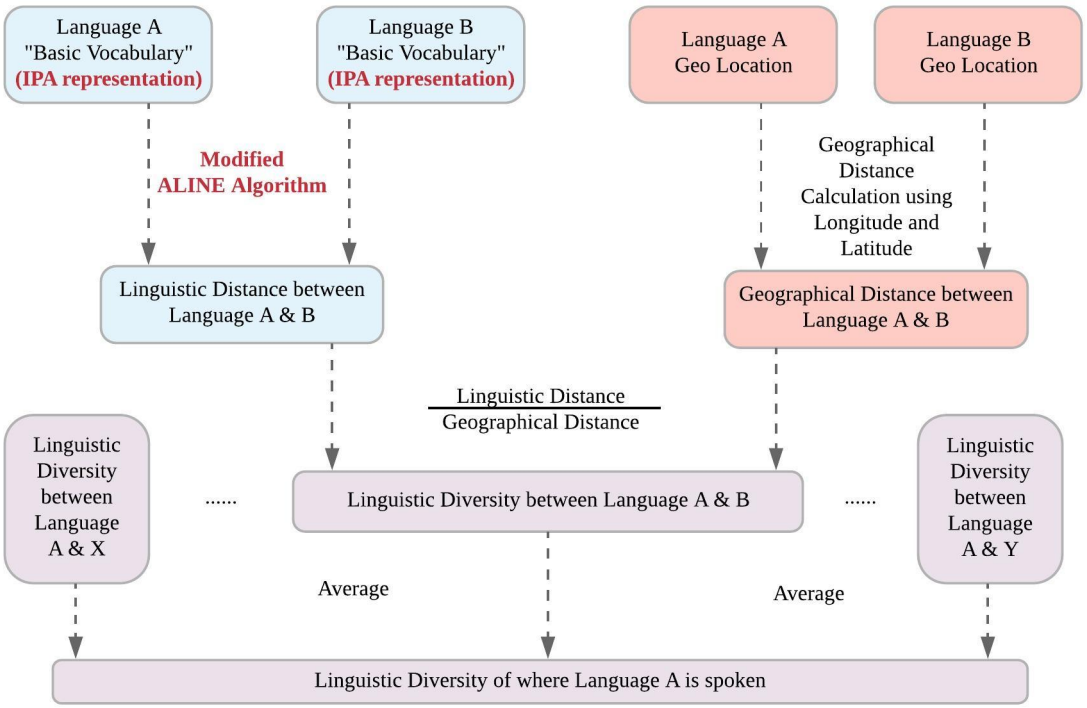



Figure 5.1: Block diagram of the modified ASJP model proposed by this thesis. The modifications are marked as red.

For linguistic distance calculation, the modified model uses words in the Swadesh list as data and processes the data in a fashion similar to the ALINE algorithm. All

the words are encoded in the International Phonetic Alphabet (IPA).

The model starts by converting the words to a list of features. For example, /b/ is a voiced bilabial stop that has the default tone and length, and is unnasalized and unlateralized. Each feature can be represented by a numeric value, as shown in Table 5.1. In that case, /b/ would corresponds to such a set of feature values: {Place: 100, Manner: 100, Tone: 50, Voice: 100, Length: 33, Nasal: 0, Lateral: 0}.

| Feature Name | Value Name | Value |
|---|---|---|
| Place (consonant) | Bilabial | 100 |
| | Labiodental | 95 |
| | Dental | 90 |
| | Alveolar | 85 |
| | Postalveolar | 80 |
| | Retroflex | 75 |
| | Alveolo-palatal | 75 |
| | Palato-alveolar | 70 |
| | Palatal | 60 |
| | Velar | 50 |
| | Uvular | 40 |
| | Pharyngeal | 20 |
| | Glottal | 0 |
| Place (vowel) | Front | 100 |
| | Central | 50 |
| | Back | 0 |
| Manner (consonant) | Stop | 100 |
| | Trill | 80 |
| | Tap | 70 |
| | Affricate | 60 |
| | Fricative | 50 |
| | Approximant | 30 |
| Manner (vowel) | High-vowel | 20 |
| | Mid-vowel | 10 |
| | Low-vowel | 0 |

| Feature Name | Value Name | Value |
|---|---|---|
| Tone | High | 100 |
| | Rising | 75 |
| | Mid (default) | 50 |
| | Falling | 25 |
| | Low | 0 |
| Voice | Voiced | 100 |
| | Ejective | 75 |
| | Voiceless | 50 |
| | Breathy | 25 |
| | Aspirated | 0 |
| Length | Long | 100 |
| | Half-long | 66 |
| | Default | 33 |
| | Extra-short | 0 |
| Round (vowel) | Rounded | 100 |
| | Unrounded | 0 |
| Nasal | Nasalized | 100 |
| | Unnasalized | 0 |
| Lateral | Lateralized | 100 |
| | Unlateralized | 0 |

Table 5.1: Feature values of the modified model. The numeric values are chosen arbitrarily based on *a priori* linguistic knowledge.

| Feature Name | Salience |
|---|---|
| Place | 50 |
| Manner | 50 |
| Tone | 10 |
| Voice | 10 |
| Length | 10 |
| Round | 10 |
| Nasal | 5 |
| Lateral | 5 |

| Operation | Cost |
|---|---|
| Skip | -1000 |
| Substitution | 3500 |
| Expansion | 4500 |

Table 5.2: Feature salience and word transformation operation cost of the modified model. The numeric values are chosen arbitrarily based on *a priori* linguistic knowledge and might need further tuning.

Any two words should be able to transform to each other through a set of word transformation operations. A word transformation operation can be insertion, deletion, substitution of a phoneme, breaking one phoneme into two, or coalescence of two phonemes into one. For each word transformation operation, the two phonemes involved might have different values for some features, and the similarity score between them is calculated using the following formula. The feature saliences and operation costs are shown in Table 5.2.

maybe reformat this equation. putting the summation part on a new line might help?

$$\text{operation similarity} = \text{operation cost} - \sum_{\text{feature}} (|\text{phoneme1 feature value} - \text{phoneme2 feature value}| * \text{feature salience})$$

The absolute value symbol is not very clear on a first read.

The similarity between two words is the sum of the similarity scores of all the operations used to transform them to each other. If two words can be transformed through multiple combinations of operations, the similarity between them is the highest possible similarity score calculated from the several sets of operations.

same here

$$\begin{aligned}\text{word similarity} &= max(\sum_{\text{operation}} \text{operation similarity}) \\ &= max(\sum_{\text{operation}} (\text{operation cost} - \sum_{\text{feature}} (|\text{word1 feature value} - \\ &\quad \text{word2 feature value}| * \text{feature salience})))\end{aligned}$$

In this manner, the modified model computes the phonological similarity between each pair of semantically equivalent words through dynamic programming (Wagner and Fischer, 1974). Figure 5.2 exhibits the feature comparison between 'bond'/bɑnd/ and 'bat'/bæt/, which results in the similarity matrix Table 5.3.

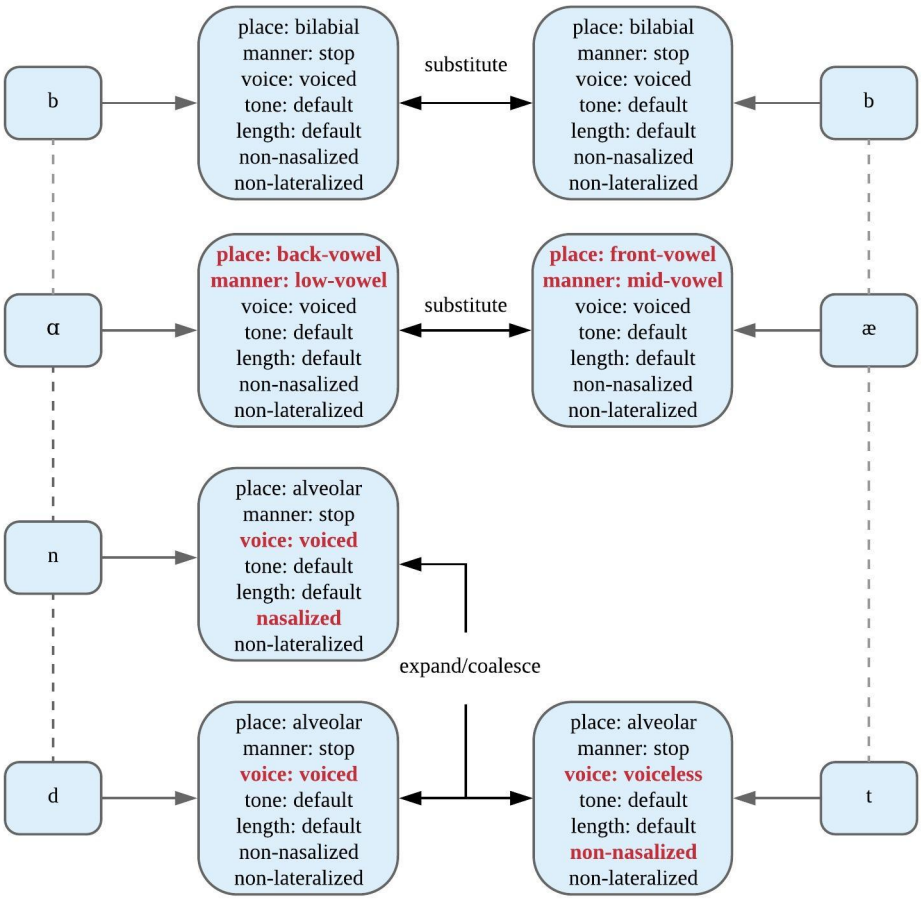



Figure 5.2: Feature comparison between 'bond'/bɑnd/ and 'bat'/bæt/. The different features are marked as red.

| | | b | ɑ | n | d |
|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 |
| b | 0 | 3500 | 2500 | 2250 | 2750 |
| æ | 0 | 2500 | 1500 | 1250 | 1750 |
| t | 0 | 2250 | 1250 | 4000 | 4500 |

Table 5.3: An example of similarity matrix of the modified model.

Let the value in the $i$th column and $j$th row of the similarity matrix be $S(i, j)$. $S(i, j)$ is the similarity between the prefix of word1 of length $i$ and the prefix of word2 of length $j$. For example, in Table 5.3, 1500 is the similarity between "bɑ" and "bæ." $S(i, j)$ is calculated through dynamic programming with the following recurrence relation. If either $i$ or $j$ is 0, then $S(i, j)$ is 0, as empty strings are prescribed as having no resemblance to any strings, including themselves. If both

$i$ and $j$ are not 0, then the relation between $\text{word1}_i$ and $\text{word2}_j$ can be one of the five word transformation operations. The highest possible similarity score generated by the five operations will be recorded in the similarity matrix. After filling in the whole matrix, the value in the right bottom cell will be the similarity between the two words.

$$S(i,j) = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ \max(S(i-1,j) + \text{skip}(\text{word1}_i), & \\ \quad S(i,j-1) + \text{skip}(\text{word2}_j), & \\ \quad S(i-1,j-1) + \text{substitute}(\text{word1}_i, \text{word2}_j), & \text{otherwise} \\ \quad S(i-2,j-1) + \text{expand}(\text{word1}_{i-1,i}, \text{word2}_j), & \\ \quad S(i-1,j-2) + \text{expand}(\text{word1}_i, \text{word2}_{j-1,j})) & \end{cases}$$

good!

As introduced above, the method of measuring word similarity in the modified model is in general similar to the ALINE algorithm, yet it differs from the original ALINE algorithm in the following aspects:

1. Both the ALINE algorithm and the modified model uses phonetic transcriptions of words as raw data. However, the ALINE algorithm encodes the data with the English alphabet, while the modified model uses the International Phonetic Alphabet (IPA). Compared to the English alphabet, the IPA has more symbols and can represent a wider variety of phonological differences.

2. Because the modified model encodes its data in the more sophisticated IPA, it contains more categories of features and more features in some categories than the original ALINE algorithm. The feature values and feature saliences are also different.

Part 1 and 2 seem connected. Want to combine them?

3. The vowel operation cost is removed, since it acts as a remedy of the poorly-designed feature categories in the original ALINE algorithm.

4. While the ALINE algorithm chooses the highest value in the similarity matrix as the overall similarity between two words, the modified model picks the right bottom value. The ALINE algorithm is initially designed to detect the parts in two words that resemble the most, but the modified model should always compare two complete words from the Swadesh list.

The linguistic similarity between two semantically equivalent words is converted to a distance using the following formula. The similarity is first normalized with respect to the self-similarity of word1 and word2, then subtracted from 1 to be the

linguistic distance between the two words. (The self-similarity is the similarity between a word and itself.) By doing this, the same phonological difference is weighted less in the linguistic distance measurement for longer, more complicated phonetic sequences.

$$d(\text{word1, word2}) = 1 - \frac{2 * \text{similarity(word1, word2)}}{\text{similarity(word1, word1)} + \text{similarity(word2, word2)}}$$

Then, the linguistic distance between two languages is obtained through averaging the linguistic distances between pairs of semantically equivalent words. The ratio between the linguistic distance and the geographical distance is the linguistic diversity of the region that the two languages span.

Finally, the modified model locates the language family homeland by picking the most diverse location among the homeland candidates. The model treats the speaking regions of the current languages in the family as the homeland candidates. For each current language, the model averages the diversity between it and other languages to be the diversity of its speaking region. The more diverse a homeland candidate is, the more likely it is the homeland of the language family.

The modified model is implemented in Python 2.7. The program and data is open-source and stored online in a git repository. [1]

# 6 Sample Input and Output

Following are some sample input and output of the modified model. Suppose there is a language family that consists of only three languages, French, German and Russian. Using a 5-word Swadesh list for each language, the modified model will measure the linguistic and geographical distance between them and hypothesizes the homeland location of the family. Table 6.1 is the sample input.

| Meaning | French | German | Russian |
|---|---|---|---|
| 'mountain' | mɔ̃taɲ | bɛrk | gɔˈra |
| 'night' | nɥi | naxt | nɔtʃ |
| 'full' | plɛ̃ | fɔl | ˈpɔɬnɨj |
| 'new' | nuvo | nɔ͡y | ˈnɔvɨj |
| 'name' | nɔ̃ | ˈnaːmə | ˈimʲa |

Table 6.1: Sample input encoded in the IPA.

[1] https://github.com/ZitingShen/ImprovedASJP

Table 6.2 shows the similarity between semantically equivalent words of different languages. High similarity scores indicate that the vocabulary in the two languages are similar in pronunciation.

| Meaning | French and German | French and Russian | German and Russian |
|---|---|---|---|
| 'mountain' | 6250 | 8000 | 2750 |
| 'night' | 3000 | 3500 | 3500 |
| 'full' | 2250 | 3100 | 3850 |
| 'new' | 7250 | 8500 | 6750 |
| 'name' | 2750 | 250 | 3330 |

Table 6.2: The similarity between semantically equivalent words from different languages in the sample input.

From Table 6.2, the distance between semantically equivalent words can be calculated, as shown in Table 6.3. Low distance scores indicate that the two languages are similar in pronunciation.

| Meaning | French and German | French and Russian | German and Russian |
|---|---|---|---|
| 'mountain' | 0.6032 | 0.4921 | 0.8036 |
| 'night' | 0.7551 | 0.7143 | 0.75 |
| 'full' | 0.7857 | 0.8032 | 0.7556 |
| 'new' | 0.4082 | 0.4604 | 0.5179 |
| 'name' | 0.7381 | 0.9714 | 0.7282 |

Table 6.3: The distance between semantically equivalent words from different languages in the sample input.

Based Table 6.3, the linguistic distance between the three languages can be measured. Meanwhile, the geographical distance between the languages can be computed from the language geographical locations provided in the ASJP database. Figure 6.1 exhibits the linguistic and geographical distances between French, German, and Russian. Low linguistic distance scores indicate that the two languages are similar in pronunciation.

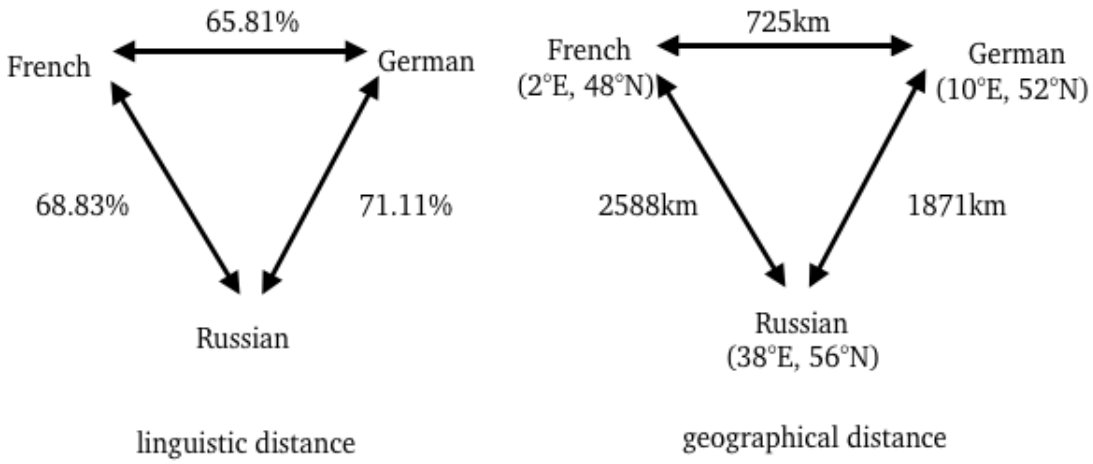



Figure 6.1: The linguistic and geographical distances between French, German, and Russian.

The linguistic diversity of the regions where French, German, and Russian are respectively spoken is calculated as below. Higher diversity values indicate that the region where the language is spoken is relatively more linguistically diverse.

$$\text{Diversity}_{\text{French}} = (\frac{0.6581}{725} + \frac{0.6883}{2588}) \div 2 = 0.0005868$$
$$\text{Diversity}_{\text{German}} = (\frac{0.6581}{725} + \frac{0.7111}{1871}) \div 2 = 0.0006439$$
$$\text{Diversity}_{\text{Russian}} = (\frac{0.6883}{2588} + \frac{0.7111}{1871}) \div 2 = 0.0003230$$

Because the geographical region where German is spoken has the highest linguistic diversity, the modified model will propose the homeland of the language family to be near where German is spoken now.

# 7 Testing and Result

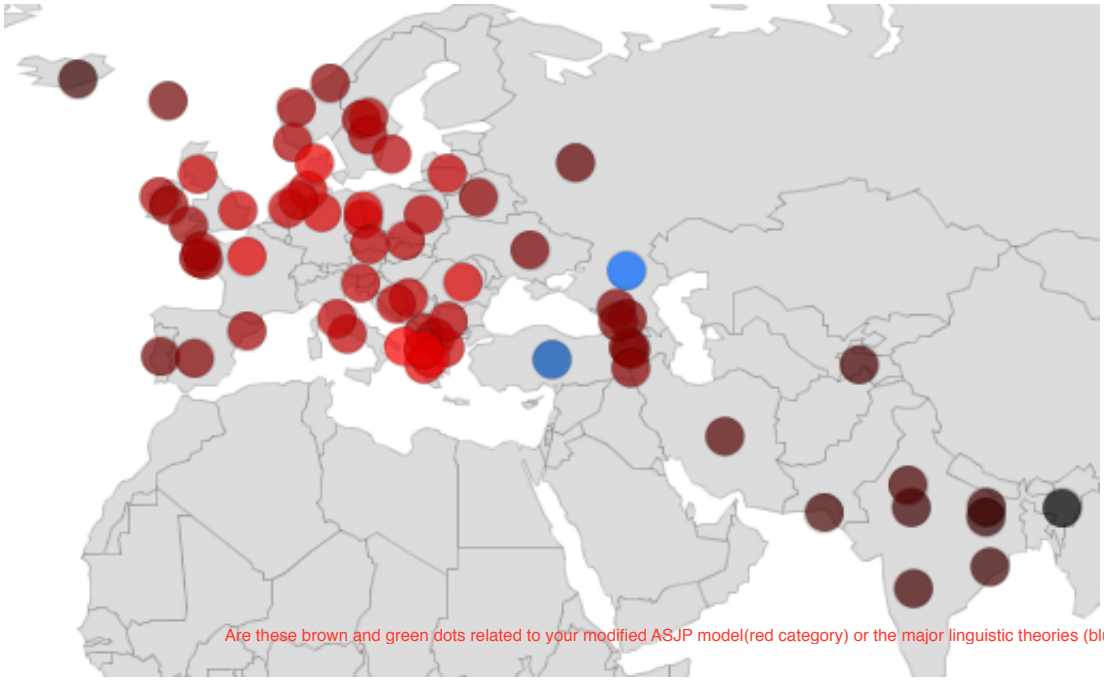
The modified ASJP model is tested on the Indo-European family, and the results are compared to current major theories of the Indo-European homelands.

## 7.1 Data

The modified model can be tested on the Indo-European family with data from the Indo-European Lexical Cognacy Database (IELex). The IELex database is a

combination of Comparative Indo-European Database and the Computational Phylogenetics in Historical Linguistics project, both of which have been used in prominent previous works (Bouckaert et al., 2012; Dunn et al., 2011). This database contains the 200-item Swadesh lists encoded in orthographic and phonological forms (i.e. encoded in IPA) for 163 languages, although most of them have some words missing. After filtering out those languages without sufficient words, we obtain 66 languages with over 40 words encoded in IPA, which is about one fifth of existing and extinct Indo-European languages. Testing the model with only 66 languages rather than the complete Indo-European language family can possibly bias the results.

## 7.2 Result



Are these brown and green dots related to your modified ASJP model(red category) or the major linguistic theories (blue category)?

maybe make it bold to distinguish it from your long description

Do you mean a candidate language?

Figure 7.1: Results of the modified ASJP model. Each dot represents the geographical location of an Indo-European language and a candidate of the homeland. Redder dots are more linguistically diverse and are more likely to be the Indo-European homeland. In contrast, blue dots are the Indo-European homelands proposed by major linguistic theories.

The figure description is pretty long!
It would be nice if you could provide a scale to describe the notation used in the figure. And leave the more detailed explanation later.

(It would be better if I can provide result visualization of the original ASJP model. However, cleaning up the data is hard, and I'm still working on it.)

I think it's actually okay if you didn't have that visualization as long as you can describe in more detail what's the differences between your ategorization and those from linguistic theories

Shown in Figure 7.1, the modified ASJP model proposes that the Balkans is the homeland of the Indo-European family. Such result accurately reflects that in practice

the Balkans is one of the most linguistically diverse region within the Indo-European family. However, it does not align with the two major linguistic theories about Indo-European homelands. As shown in Figure 7.1, the two homeland locations proposed by the major linguistic theories are more to the east compared to the one proposed by the improved ASJP model. What might be reason of this discrepancy?

By observing Figure 7.1, it is obvious that the IELex dataset itself is moderately biased and leads to the inaccurate homeland identification result. The dataset is too Euro-centric, i.e. it only contains a few Indo-Iranian languages which are mainly spoken in South, Central and Western Asia. However, in reality, more than one third of Indo-European languages are Indo-Iranian. Without enough data points about the Indo-Iranian languages, the proposed homeland is likely to be biased and shifted westward. Make it more clear that this paragraph is talking about the linguistic theories. You believe that it might leads to the improper classifications

I am a little confused about the transition here

When analyzing the results, we also realize that there is an urgent need to visualize the data points and results of computational models used for language family homeland identification. Without visualization, it is hard to learn about some prominent bias in the dataset and filter out the abnormal data points. Visualization does not only make result examination more intuitive and perceptive, but also prompts inspiring thoughts for further modification and refinement of the existing models.

remove "does" and change make to "makes"?

# 8 Future Work

Maybe incorporate/recall what's currently done to clarify what each future work is related to

Future improvements of the model proposed in this thesis should be focused on fixing the problematic assumptions of the original ASJP model. For linguistic distance measurement, syntax and morphological attributes could be taken into account in addition to phonological distances to thoroughly measure the linguistic distance. Additionally, the accuracy of the modified model might be limited by choosing the speaking regions of the current languages to be the homeland candidates. The possibility that no current languages might be spoken spoken in the homeland should not be ignored, thus it is worth experimenting on handpicking homeland candidate locations and comparing with the current model.

Moreover, the center of gravity principle employed by the ASJP model assumes most language changes are innovations generated in the evolution of the ancient languages, and fails to recognize some language changes results from the influence of neighboring language families (Mallory and Adams, 2006, p. 445-446). It would be ideal to separate these two kind of language changes. One way to achieve this is to separate cognates, i.e. "innovations" in language evolution, from borrowed lexical

items in the original data. Nevertheless, such separation might require a fair amount of linguistic expertise and labor. Another simpler way might be to subtract the diversity between languages in this language family and languages in other language families from the total diversity for each candidate. The result could be considered as the diversity generated by genetic relationships only.

Other possible improvements include examining the evolution of sounds with respect to possible migration routes of the language speakers. Because sound changes usually happen during the migration of language speakers, a sound might consistently evolved in the languages spoken in a geographical area. Such consistent change of sound might indicate the direction of migration of ancient language speakers and the location of language family homeland.

With respect to testing, it would be ideal to test the modified model on another language family to see how well the model can handle different kinds of complexities in language family homeland identification.

# 9 Conclusion

This paper has presented a modification of the ASJP model to incorporate more phonological details in language similarity and distance measurement to improve the accuracy of automatic homeland identification. The results are reasonable yet biased due to the bias in the model assumptions and testing data. Apart from the biased dataset, the modified ASJP model does provide a sensible estimation about the most linguistically diverse region within the language family. Whether such region should be recognized as the homeland of a language family needs more discussion in the field of linguistics.

clear conclusion

The modified ASJP model s based on the ASJP model and the ALINE algorithm (Kondrak and Hirst, 2002; Wichmann et al., 2010a). It does its best efforts to identify the language family homeland accurately without adding complexity to the methodology or demanding fine manual preprocessing of the dataset. It treats the phonological differences between the Swadesh lists of two languages as the the linguistic distance between them, computes the ratio between their linguistic distance and their geographical distance as the diversity of the region that the two languages span, obtains the aggregate diversity of candidate homeland locations, and ranks the aggregate diversity of the candidate locations as their likelihood to be the homeland. Such transparent methodology and simple requirements of data of the modified model render it universally applicable to the world's language families. For language families

that have special characteristics, the modified model can also be modified to adapt to the characteristics.

In addition to the integration of an improved ALINE algorithm proposed in this thesis, the ASJP model has a lot of other potential to be improved. With the limit of linguistic expertise, labor, and time, and the increasing amount and availability of data, computational approaches on linguistic topics like language family homeland identification is highly valuable and worth more exploration in the future.

# 10 Acknowledgements

I wish to express my sincere gratitude to Professor Deepak Kumar, Professor Dianna Xu and Professor Jonathan Washington to supervise this thesis and and firmly support my academic interests in my college years. I also want to thank Xuan Huang, Nicole Petrozzo, Xinyue Zhang and Tu Luan for their valuable feedback for the thesis drafts.

I am very grateful to the consistent guidance from the computer science and linguistics departments of Bryn Mawr College, Haverford College, and Swarthmore College. My educational experience here has promoted me to higher ground and permanently changed my life.

Finally, I wholeheartedly appreciate the company from my close friends and my mom. Without their selfless assistance I would not have been able making through the ups and downs in my college years. Special thanks goes to Rachel Xu, my Bryn Mawr buddy since 2014.

# References

Anthony, David W. and Don Ringe. 2015. The Indo-European Homeland from Linguistic and Archaeological Perspectives. *Annual Review of Linguistics* 1(1). 199–219.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant and Eric W. Holman. 2009. Adding Typology to Lexicostatistics: A Combined Approach to Language Classification. *Linguistic Typology* 13(1).

Bergsland, Knut and Hans Vogt. 1962. On the Validity of Glottochronology. *Current Anthropology* 3(2). 115–153.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard and Q. D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 337(6097). 957–960.

Brown, Cecil H., Eric W. Holman, Søren Wichmann and Viveka Velupillai. 2008. Automated Classification of the World's Languages: A Description of the Method and Preliminary Results. *Language Typology and Universals* 61(4). 285–308.

Downey, Sean S., Brian Hallmark, Murray P. Cox, Peter Norquest and J. Stephen Lansing. 2008. Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction. *Journal of Quantitative Linguistics* 15(4). 340–369.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson and Russell D. Gray. 2011. Evolved Structure of Language Shows Lineage-specific Trends in Word-order Universals. *Nature* 473(7345). 79–82.

Ethnologue. 2017. Ethnologue: Languages of the World. Tech. rep.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature* 426(6965). 435–439.

Heggarty, Paul. 2000. Quantifying Change Over Time in Phonetics. In Colin Renfrew, April M. S. McMahon and R. L. Trask (eds.), *Time Depth in Historical Linguistics*, Papers in the Prehistory of Languages, 531–562. Cambridge: McDonald Institute for Archaeological Research. OCLC: ocm46657163.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller and Dik Bakker. 2008. Explorations in Automated Language Classification. *Folia Linguistica* 42(3-4).

Huff, Paul and Lonsdale. 2011. Positing Language Relationships Using ALINE. *Language Dynamics and Change* 1(1). 128–162.

Kondrak, Grzegorz. 2000. A New Algorithm for the Alignment of Phonetic Sequences. *NAACL 2000 Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* 288–295.

Kondrak, Grzegorz and Graeme Hirst. 2002. *Algorithms for Language Reconstruction*, vol. 63. Toronto: University of Toronto.

Levenshtein, Vladimir I. 1966. Binary Codes Capable of Deletions, Insertions, and Reversals. *Soviet Physics-Doklady* 10(8). 707–710.

Mallory, J. P. and Douglas Q. Adams. 2006. *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World*. New York: Oxford University Press.

Millar, Robert McColl and R. L. Trask (eds.). 2015. *Trask's Historical Linguistics*. New York: Routledge, 3rd edn.

Onions, Charles T. (ed.). 1996. *The Oxford Dictionary of English Etymology*. Oxford: Oxford Univ. Press, repr edn. OCLC: 246698002.

Swadesh, Morris. 1950. Salish Internal Relationships. *International Journal of American Linguistics* 16(4). 157–167.

Wagner, Robert A. and Michael J. Fischer. 1974. The String-to-String Correction Problem. *Journal of the ACM* 21(1). 168–173.

Wichmann, Søren, Eric W. Holman, Dik Bakker and Cecil H. Brown. 2010a. Evaluating Linguistic Distance Measures. *Physica A: Statistical Mechanics and its Applications* 389(17). 3632–3639.

Wichmann, Søren, André Müller and Viveka Velupillai. 2010b. Homelands of the

World's Language Families: A Quantitative Approach. *Diachronica* 27(2). 247–276.